applied sciences

MDPI

Article

SEM-Net: A Social–Emotional Music Classification Model for Emotion Regulation and Music Literacy in Individuals with Special Needs

Yu-Chi Chou [1,*], Shan-Ken Chien [1], Pen-Chiang Chao [2], Yuan-Jin Lin [3], Chih-Yun Chen [4], Kuang-Kai Yeh [5], Yen-Chia Peng [5], Chen-Hao Tsao [4], Shih-Lun Chen [4] and Kuo-Chen Li [6]

1 College of Humanities and Education, Chung Yuan Christian University, Taoyuan City 32023, Taiwan; shanken@cycu.edu.tw
2 Department of Special Education, National Taiwan Normal University, Taipei City 106308, Taiwan; chaopc@ntnu.edu.tw
3 Department of Program on Semiconductor Manufacturing Technology, Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan City 701401, Taiwan; m28121562@gs.ncku.edu.tw
4 Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan City 32023, Taiwan; g11376049@cycu.edu.tw (C.-Y.C.); s11126361@cycu.edu.tw (C.-H.T.); chrischen@cycu.edu.tw (S.-L.C.)
5 Department of Information and Computer Engineering, Chung Yuan Christian University, Taoyuan City 32023, Taiwan; s11227128@cycu.edu.tw (K.-K.Y.); s11227130@cycu.edu.tw (Y.-C.P.)
6 Department of Information Management, Chung Yuan Christian University, Taoyuan City 320317, Taiwan; kuochen@cycu.edu.tw
* Correspondence: yuchichou@cycu.edu.tw

Featured Application: We developed a social–emotional music classification model named SEM-Net specifically designed for individuals with special needs, achieving a final accuracy of 94.13%, outperforming existing models by at least 27%.

Abstract: This study aims to establish an innovative AI-based social–emotional music classification model named SEM-Net, specifically designed to integrate three core positive social–emotional elements—positive outlook, empathy, and problem-solving—into classical music, facilitating accurate emotional classification of musical excerpts related to emotional states. SEM-Net employs a convolutional neural network (CNN) architecture composed of 17 meticulously structured layers to capture complex emotional and musical features effectively. To further enhance the precision and robustness of the classification system, advanced social–emotional music feature preprocessing and sophisticated feature extraction techniques were developed, significantly improving the model's predictive performance. Experimental results demonstrate that SEM-Net achieves an impressive final classification accuracy of 94.13%, substantially surpassing the baseline method by 54.78% and outperforming other widely used deep learning architectures, including conventional CNN, LSTM, and Transformer models, by at least 27%. The proposed SEM-Net system facilitates emotional regulation and meaningfully enhances emotional and musical literacy, social communication skills, and overall quality of life for individuals with special needs, offering a practical, scalable, and accessible tool that contributes significantly to personalized emotional growth and social–emotional learning.

Keywords: convolutional neural network; feature extraction; music classification; social–emotional learning; social regulation

check for updates

Academic Editors: Miguel Angel Patricio and Luis Usero Aragonés

Received: 18 March 2025
Revised: 9 April 2025
Accepted: 9 April 2025
Published: 10 April 2025

Citation: Chou, Y.-C.; Chien, S.-K.; Chao, P.-C.; Lin, Y.-J.; Chen, C.-Y.; Yeh, K.-K.; Peng, Y.-C.; Tsao, C.-H.; Chen, S.-L.; Li, K.-C. SEM-Net: A Social–Emotional Music Classification Model for Emotion Regulation and Music Literacy in Individuals with Special Needs. *Appl. Sci.* **2025**, *15*, 4191. https://doi.org/10.3390/app15084191

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Social–emotional functioning includes socio-emotional cognition and social behaviors [1]. Positive social–emotional functioning involves understanding and regulating emotions and building positive relationships. It also includes a couple of essential components. One essential component, positive outlook, refers to maintaining a positive attitude toward events and circumstances, clearly recognizing one's psychological state and abilities, and proactively engaging in social interactions. Research indicates that a positive outlook contributes significantly to emotional resilience when facing adversity and is considered an essential life skill [2]. Additionally, practical social problem-solving skills contribute to positive feedback and recognition in social interactions, enhancing psychological resilience, self-efficacy, and social adaptation [3,4]. The other essential component contributing to positive social–emotional functioning involves empathy, the ability to understand and infer others' emotions, perspectives, and thoughts and respond appropriately based on observed or imagined emotional experiences. A critical aspect of empathy is perspective-taking, which allows individuals to adopt others' viewpoints, facilitating deeper comprehension of social contexts and emotional states [5,6].

Musical literacy and emotion share a close and interactive relationship. Musical literacy refers to an individual's ability to understand, appreciate, and create music, while emotion encompasses responses to external stimuli. Unlike language literacy, musical literacy requires continuous listening and experiential learning [7]. Musical literacy enhances music appreciation and plays a crucial role in emotion regulation, communication, and social interactions. Integrating instrumental performance with cognitive and emotional processing activities facilitates problem-solving and attribution retraining [8]. The duration of music education interventions significantly affects their effectiveness, with longer interventions generally leading to more pronounced improvements in emotion regulation, benefiting both general and high-risk populations [9,10]. For example, children listening to soft music-accompanied storytelling in teaching self-regulation of emotions can better recognize their inner emotions and thoughts. Through guided discussions, they learn how story characters manage emotional changes, enhancing social communication and emotional behavior skills [11]. This study explores the impact of classical music on social–emotional development. Classical music holds value due to its diverse cultural origins, historical significance, and intellectual depth. It uniquely reflects the richness of human experiences and effectively stimulates imagination, unlike many popular music genres constrained by predetermined lyrical frameworks [12]. Numerous studies have demonstrated that classical music can directly influence emotional states [13–15]. For example, Baumgartner et al. [16] studied how visual and musical stimuli affect brain processing. Using emotionally intense images from the International Affective Picture System and classical music clips, they induced happiness, sadness, and fear. Physiological data, including heart rate, skin conductance, respiration, temperature, and self-reported emotions, were collected. Results showed that combining visual and musical stimuli most accurately reflected emotional experiences.

Music genre classification (MGC) is a task within the Music Information Retrieval (MIR) domain to achieve a computational understanding of musical semantics. For a given song, the classifier predicts relevant musical attributes and features. Based on the task definition, almost infinite classification objectives range from genres, emotions, and instruments to broader concepts. These features are crucial markers that distinguish various musical styles. For instance, string instruments in classical music and synthesizers in pop music directly reflect their stylistic characteristics. Arrangements and instrumentation alter audio's spectral distribution and sonic characteristics, providing critical inputs for machine learning models. These inputs enhance the accuracy of music classification,

addressing diverse needs such as recommendation systems, music management, and emotion analysis [17]. However, the process of music classification should primarily rely on quantitative descriptors of the purely musical aspects of a piece. A. Anglade et al. [18] utilized a novel genre classification framework for low-level signal-based and high-level harmonic features, covering categories like classical, jazz, and pop genres. Similarly, C. Pérez-Sancho et al. [19] introduced a symbolic audio-based music genre classification framework that categorizes music by analyzing harmonic structures. Regarding deep learning techniques, standard classification models employ convolutional neural networks (CNNs). These models incorporate preprocessing and annotation techniques for input sources, balancing label quality and data volume. This approach addresses the limitations of most existing Sound Event Recognition (SER) datasets, which are often relatively small or domain-specific [20].

In fully automated audio analysis and music feature extraction, Mel-Frequency Cepstral Coefficients (MFCC) play a pivotal role as a feature extraction method. MFCC transforms audio signals into a spectral cepstral representation, capturing the audio's frequency and energy distribution characteristics. Modern research often leverages deep learning models, such as DNN, CNN, and LSTM, using MFCC as input features to improve music classification performance [21]. Apart from MFCC, Chroma features are essential for describing melodic and harmonic characteristics in music by mapping audio data into 12 semitone-based pitch categories. Chroma features capture the fundamental aspects of melody, harmony, and tonality. F. Zalkow et al. [22] provide a novel application of Chroma features introduced for music retrieval based on scores and audio. The approach integrates Connectionist Temporal Classification (CTC) loss for training, reducing reliance on strongly aligned training data. Traditional methods typically require precise alignment between scores and audio. Still, CTC loss only necessitates weak annotations (e.g., start and end times of musical themes), significantly simplifying data preparation and improving efficiency. Among various frequency domain analysis techniques, F.A.D. Rí et al. [23] explored the simultaneous use of MFCC, Chroma, and Spectral Contrast for Sound Event Recognition (SER). By implementing CNNs with convolutional attention blocks, the study achieved an average classification accuracy of 82.5%. Additionally, using five-fold cross-validation, the highest accuracy reached 92.71%, demonstrating enhanced generalization capability in feature extraction.

Based on theoretical knowledge, the AI social–emotional music model focused on integrating positive emotional elements into classical music. This study aimed to develop an innovative AI model for improving emotion regulation and music literacy for individuals with special needs, including those with disabilities, such as intellectual and developmental disabilities. Since individuals with disabilities may have a higher likelihood of exhibiting social and emotional challenges, including internalizing and externalizing behavioral problems, we propose that implementing an emotional music model could be an effective mechanism for coping with social and emotional conflicts. Research questions included: (1) feature extraction of audio regions of interest in classical music; (2) development of deep learning models applicable to three different social–emotional contexts; and (3) model validation techniques and comparison with current research.

2. Materials and Methods

This section is divided into four subsections. The first subsection describes the three types of social–emotional music studied in this research and outlines the extraction of three corresponding musical features. The second subsection focuses on multi-feature fusion through spectral characteristics to enhance the accuracy of CNN training for the three music categories. The third subsection extracts the features of the music after pre-

processing. The fourth subsection evaluates and validates the training results using the proposed music emotion classification model. Figure 1 illustrates the overall methodology, comprising four main stages: data processing, feature extraction, model training, and performance evaluation.

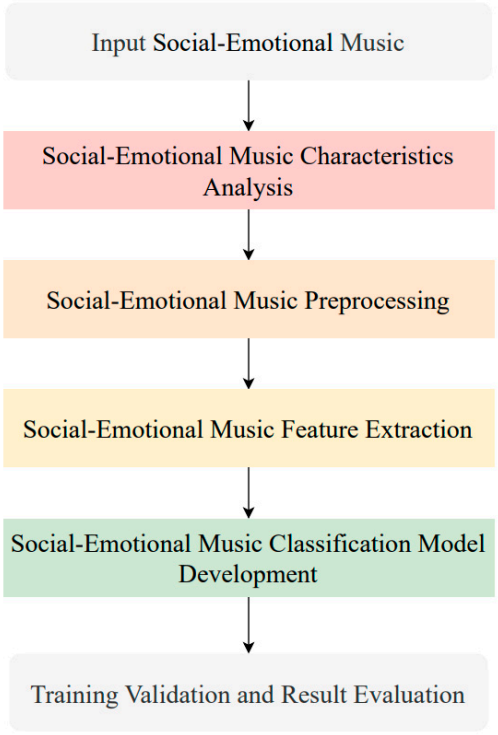


Figure 1. Social–emotional music classification flow chart.

2.1. Social–Emotional Music Characteristics Analysis

This study employs an AI model to link three positive emotional elements—outlook, empathetic perspective-taking, and problem-solving—with corresponding musical elements. A special education team defined and interpreted these emotional elements, constructing relevant social contexts and examples and developing strategies to enhance these emotional capabilities. The first element, "outlook", is musically represented by melodic fluency, lyrical coherence and stability, and thematic richness and contrast, mirroring the pleasant, meditative, and stabilizing experiences associated with New Age music. The second element, "empathetic perspective-taking", is represented by the counterpoint technique in classical music, emphasizing the interaction between melodies" note against note"—representing inter-vocal dialogue and fluidity. Individual vocal parts, while distinct, remain harmonious and non-conflicting. The third element, "problem-solving", is represented musically by the B section to A' section transition in ABA' form. This transition involves a development section (B) characterized by dramatic tension, resolving into a return to the stable initial theme (A'). After confirming the emotional strategies and musical files, a cross-disciplinary team of special education and music arts professionals conducted a bidirectional theoretical item analysis to validate the congruence between emotional and musical language.

2.2. Social–Emotional Music Pre-Processing

Differences in recording environments and background noise affect social–emotional music data, which can impact the training outcomes of deep learning models [24,25]. Therefore, music preprocessing becomes the first crucial step in enhancing the accuracy of classification models. This study applies normalization to each audio segment to stan-

dardize the amplitude range and ensure consistency in the frequency distribution across audio from different sources. Additionally, various music data augmentation techniques are employed to enhance the model's generalization ability and increase the diversity of training samples, reducing the risk of CNN model overfitting and simulating audio variations under different environments.

2.2.1. Time Stretching

Time stretching augmentation is a technique that adjusts the music speed to modify its duration. This method processes the music using a randomly selected stretch ratio, accelerating or decelerating the speed while preserving the pitch. Simulating variations in the duration of speech or music at different music speeds enables the model to adapt to temporal variations, enhancing its robustness and generalization ability in recognizing audio with speed changes. The time stretching formula is given in (1), where T_I represents the original audio duration, T_O is the modified audio duration, and R is the stretch ratio with $0.8 < R < 1.2$. When $R > 1$, the duration shortens, and when $R < 1$, the duration lengthens.

$$T_o = \frac{T_I}{R} \tag{1}$$

2.2.2. Pitch Shifting

Pitch shift augmentation is a data augmentation technique that modifies the pitch of an audio signal by applying a short-time Fourier transform (STFT) and shifting the number of semitones. In this study, the pitch is randomly altered within a range of -3 to +3 semitones, which can be raised by up to 3 semitones (+3) or lowered by up to 3 semitones (-3). This method simulates pitch variations that may occur in real-world applications while preserving the core content of the audio. By incorporating this augmentation, the model becomes more robust to pitch variations, reducing its sensitivity to pitch shifts and improving its generalization capability and classification accuracy.

2.2.3. Noise Augmentation

Noise augmentation is a technique that simulates real background noise by adding random noise to an audio signal. A Gaussian noise signal with the same length as the original audio is generated and superimposed onto the signal. Gaussian noise effectively models various real-world acoustic environments, enhancing the model's ability to generalize across different audio inputs. This augmentation improves the model's robustness in noisy conditions, aiding in audio classification tasks. The Gaussian distribution is shown in (2), where x represents the noise value, μ denotes the mean of the noise, and σ determines the noise intensity and distribution range.

$$N\left(X\middle|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

2.3. Social–Emotional Music Feature Extraction

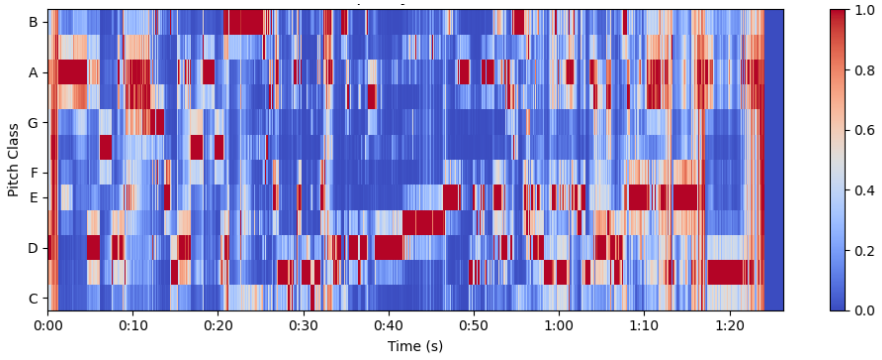
Various spectral analysis techniques are commonly employed to extract musical features in MGC pre-processing. Spectral analysis decomposes complex audio signals into different frequency components, enabling AI models to capture critical acoustic features effectively. Moreover, spectral analysis transforms time series audio signals into image or numerical representations, reducing the complexity of data processing and enhancing the AI model's ability to learn and distinguish musical patterns and differences.

2.3.1. Mel-Frequency Cepstral Coefficients Processing

MFCC is a widely used audio feature extraction technique in music classification. The core concept of MFCC is to simulate the human auditory perception of different frequencies by converting an audio signal into a set of key feature coefficients to capture its time–frequency characteristics. The MFCC computation process consists of several steps. First, the audio signal is segmented into short time frames to capture temporal variations. Next, the Fourier transform (FT) transforms the signal into the frequency domain. Then, a Mel-frequency filter bank is applied to emphasize low-frequency features while attenuating high-frequency components, aligning with human auditory perception. Subsequently, the logarithm of the filtered spectral energy is computed to compress the dynamic range. Finally, the Discrete Cosine Transform (DCT) is performed to reduce dimensionality, extracting 12 to 13 key feature coefficients, which form an audio feature vector for further classification and recognition. This study analyzed three social–emotional music elements—perspective-taking, positive outlook, and problem-solving—using Mel-frequency cepstral coefficients (MFCCs). Results indicated that the perspective-taking type exhibited smoother spectral transitions, possibly reflecting melodic continuity and stability, aligning with calm emotional characteristics. The positive outlook type demonstrated more significant variability in cepstral coefficients, potentially reflecting dynamic and expressive melodic qualities consistent with optimism and active emotional states. Lastly, the problem-solving type showed broader spectral energy distributions, likely associated with diverse emotional expressions and more dynamic musical structures.

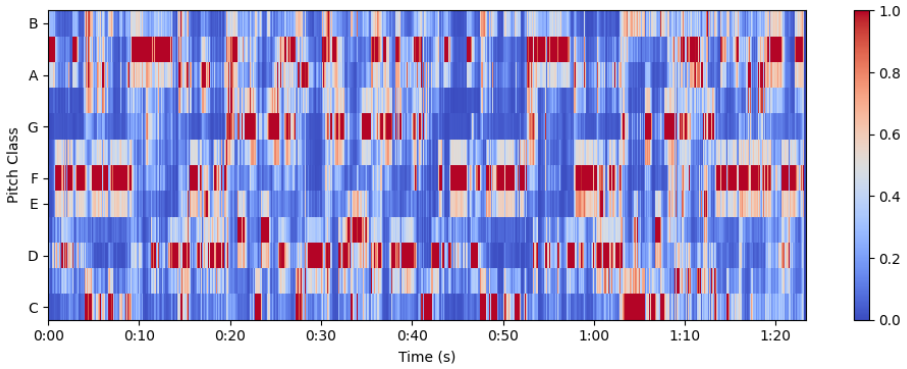
2.3.2. Chroma Processing

Chroma is a feature representation method used to describe pitch distribution by mapping pitch information from an audio signal into 12 semitone classes (C, C#, D, D#, E, F, F#, G, G#, A, A#, B) without considering absolute frequency or octave position. In other words, C4 (middle C) and C5 (high C) are mapped to the same Chroma class C. Chroma features effectively capture harmonic structures and tonal variations, making them widely used in music analysis and classification. Chroma features are utilized to analyze different categories of emotionally perceived music. Figure 2a depicts empathetic perspective-taking, where pitch energy is evenly distributed across the twelve semitone classes with relatively smooth variations, indicating a stable tonal structure and high melodic continuity. Figure 2b represents outlook, which exhibits a broader range of pitch variations and frequent key transitions, suggesting a more open melodic structure and greater musical flexibility. Figure 2c illustrates problem-solving, where specific pitch energy peaks are more prominent, indicating stronger contrast in harmonic composition and tonal transitions, which may be associated with greater emotional complexity and dynamic changes.

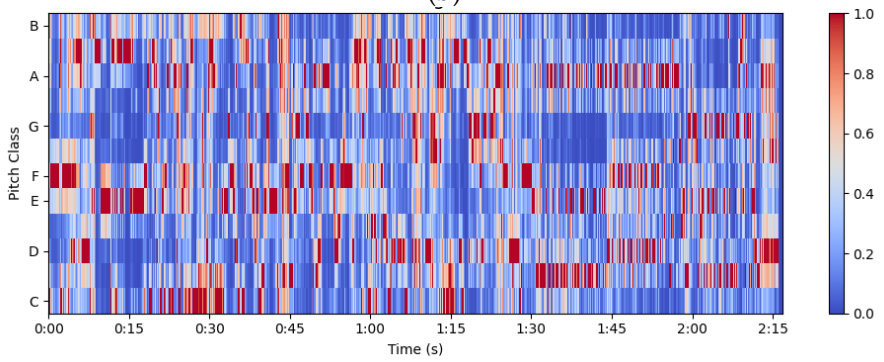


(**a**)

Figure 2. *Cont.*



(**b**)



(**c**)

Figure 2. Chroma processing results of three music categories. (**a**) Empathetic perspective-taking. (**b**) Outlook. (**c**) Problem-solving.

2.3.3. Spectral Contrast Processing

Spectral contrast is a technique used to describe audio characteristics by measuring the energy contrast between different frequency bands, focusing on the intensity difference between high and low frequencies. By analyzing the spectral peaks (high-energy regions) and valleys (low-energy regions), spectral contrast provides insights into timbre characteristics, harmonic structures, and the depth of audio layers. The calculation method is defined in (3), where $S_{max,\,b}$ represents the maximum spectral value (peak) within frequency b. In contrast, $S_{min,\,b}$ denotes the minimum spectral value (valley) within the same band.

$$C_b = 10\log_{10}\left(\frac{S_{max,\,b}}{S_{min,\,b}}\right) \tag{3}$$

This study employs spectral contrast to analyze different categories of emotionally perceived music. Figure 3a represents empathetic perspective-taking, where spectral contrast values exhibit minimal variation, indicating minor energy differences between high and low frequencies. This suggests that the timbre is relatively uniform, with smoother melodic transitions. Figure 3b illustrates the outlook, displaying a more balanced spectral contrast distribution; however, certain frequency bands still exhibit prominent energy peaks, suggesting greater tonal detail and diversity. Figure 3c represents problem-solving, demonstrating higher energy contrast, particularly in the high-frequency range, indicating more pronounced timbre variations, potentially incorporating more significant dynamic fluctuations and emotional intensity.

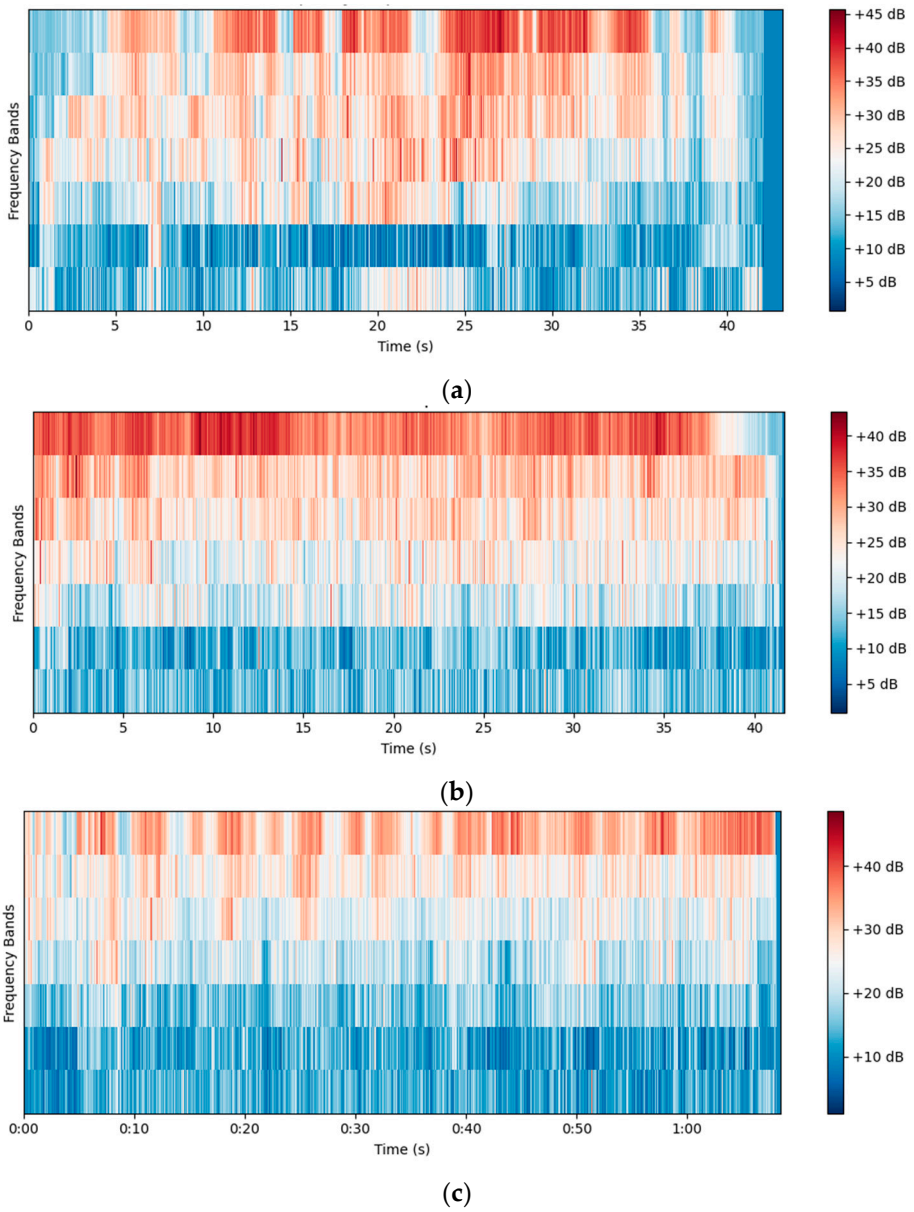


Figure 3. Spectral contrast processing results of three music categories. (**a**) Empathetic perspective-taking. (**b**) Outlook. (**c**) Problem-solving.

2.4. Social–Emotional Music Classification Neural Network Architecture

Convolutional Neural Networks (CNNs) are deep learning models well-suited for processing structured data, such as images and audio signals. By leveraging local connectivity and parameter-sharing mechanisms, CNNs can effectively extract time–frequency features from spectrogram representations while significantly reducing the number of model parameters. This study transforms music signals into spectrograms, allowing SEM-Net to learn musical styles, instrument types, and rhythm patterns. For music classification tasks, SEM-Net architecture consists of convolutional layers (to capture local time–frequency features), pooling layers (to reduce dimensionality and enhance feature invariance), and fully connected layers (to perform final classification). However, traditional CNNs may have limitations in capturing global features due to their constrained ability to model long-term dependencies in sequential data. Recent studies have explored integrating attention mechanisms (such as Squeeze-and-Excitation modules or Transformer-based self-attention) and multi-level feature fusion techniques (such as CNN-RNN hybrid models) to enhance the recognition of musical emotions and stylistic characteristics. This study

adopts a CNN-based deep learning architecture to extract time–frequency features from audio spectrograms. The complete Social–Emotional Music Classification Neural Network (SEM-Net) architecture is presented in Table 1.

Table 1. SEM-Net architecture.

Type	Activation	Filter Number/Filter Size/Stride
Image Input		Size: 250 × 200 × 3
Convolution_1	ReLU	32 Filters/5 × 5/Stride 1
Batch Normalization		Normalizes activations to stabilize training
MaxPooling_1		Pool size: 2 × 2
Dropout_1		Rate: 0.25
Convolution_2	ReLU	64 Filters/5 × 5/Stride 1
Batch Normalization		Normalizes activations to reduce internal covariate shift
MaxPooling_2		Pool size: 2 × 2
Dropout_2		Rate: 0.3
Convolution_3	ReLU	128 Filters/5 × 5/Stride 1
Batch Normalization		Normalizes activations to ensure stable gradients during backpropagation
MaxPooling_3		Pool size: 2 × 2
Dropout_3		Rate: 0.35
Flatten		Converts 2D feature maps to 1D vector for fully connected layers
Dense_1	ReLU	Units: 256
Dropout_4		Rate: 0.4
Dense_2 (Output)	Softmax	Units: Number of Classes

The detailed SEM-Net architecture is shown in Figure 4. SEM-Net receives a one-channel two-dimension input, and the input layer is denoted as I, followed by three blocks of convolutional layers: C1, C2, and C3, each with a kernel size of 5 × 5 and ReLU activation function. Each convolutional layer is followed by a batch normalization layer (B1–B3), a pooling layer (P1–P3) using 2 × 2 max-pooling, and dropout layers (D1–D3) with increasing dropout rates (0.25, 0.3, and 0.35) to prevent overfitting. After the final convolutional block, the feature maps are flattened (F) and passed through two fully connected (FC) layers. The first FC layer is dense with 256 units, followed by dropout (D4) with a rate of 0.4. The FC2 layer uses the softmax activation function to output probability scores across target music classes.

During the model training process, this study employs the cross-entropy loss function as the optimization objective, as it effectively handles multi-class classification problems, making it particularly suitable for music classification tasks. The Adam optimizer is selected with an initial learning rate of 0.001, as it combines momentum and adaptive learning rate adjustment mechanisms, enabling faster convergence compared to stochastic gradient descent (SGD) in audio classification problems. Throughout training, model performance is evaluated on the validation set at the end of each epoch. Additionally, hyperparameters such as learning rate and batch size are fine-tuned using random search. An early stopping strategy is adopted, terminating training if the validation loss does not decrease by more than 0.01 for 10 consecutive epochs to prevent overfitting. The hyperparameter configurations used in this study are shown in Table 2.

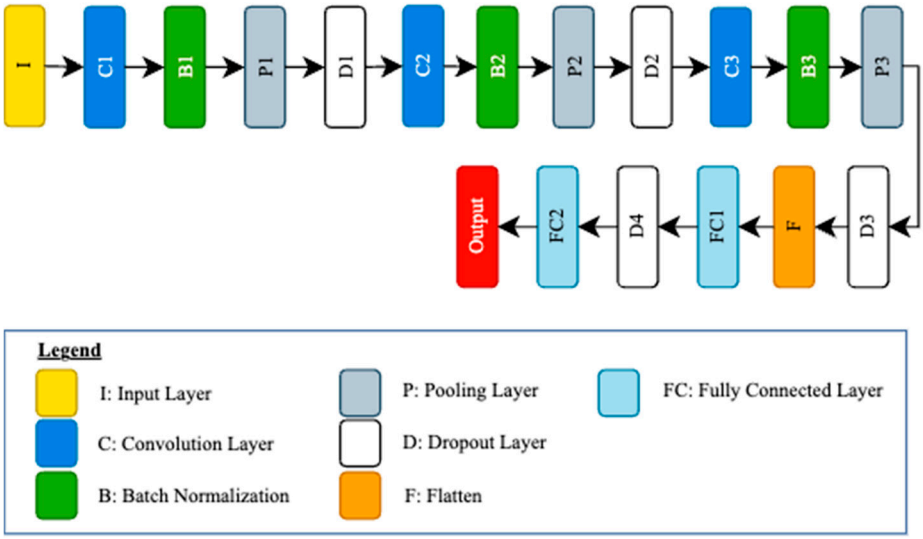


Figure 4. SEM-Net layer connection diagram.

Table 2. The HyperParameter value used in the deep learning.

Initial Learning Rate	0.001
MaxEpoch	300
Batchsize	32
Learning Rate Drop Factor	0.1
LearnRateDropPeriod	10
Shuffle	Every-epoch

The music dataset used in this study was annotated with the assistance of professional music scholars and special education experts. The size of the dataset is shown in Table 3, consisting of approximately 200 music pieces categorized into three types, each averaging around 2 min and 30 s in length. Data augmentation techniques were applied to enhance the reliability and stability of the deep learning model, including random noise addition, playback speed adjustment (ranging from 0.8× to 1.2×), and time shift modifications to the audio. The augmented dataset increased to five times the original dataset, producing 980 music pieces. The augmented dataset was then split into training and validation sets at a ratio of 8.5:1.5, with 210 music reserved for testing.

Table 3. Music dataset comparison before and after data augmentation.

Before Data Augmentation				
Set	Training Set	Validation Set	Test Set	Total
Total	130	24	42	196
After Data Augmentation				
Set	Training Set	Validation Set	Test Set	Total
Empathetic perspective-taking	250	50	75	375
Outlook	210	40	70	320
Problem-solving	190	30	65	285
Total	650	120	210	980

Accuracy, recall, precision, and mean average precision (mAP) are the key metrics for evaluating the model's performance. The calculations for these metrics are defined in (4), (5), (6), and (7). The confusion matrix is a widely used tool for evaluating the performance of classification models, particularly in multi-class classification tasks, as shown in Table 4.

It provides a matrix representation of the relationship between predicted and actual labels, allowing for an analysis of the model's performance across different classes. The confusion matrix is typically an n×n matrix, where n represents the total number of classes. Each element within the matrix indicates the correspondence between the model's predictions and the actual labels. The confusion matrix compares predicted values with actual values, presenting results in terms of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP indicates correct predictions for positive cases, TN for correct predictions of negative cases, FP for incorrect positive predictions, and FN for incorrect negative predictions. These indicators collectively assess the model's quality to ensure its effectiveness in practical applications.

$$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Fp} + \text{Tn} + \text{Fn}} \tag{4}$$

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}} \tag{5}$$

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}} \tag{6}$$

$$\text{mAP} = \frac{\sum_{q=1}^{Q} \frac{1}{n} \sum_{r \in \{0,0,1,\ldots 1\}} P_{\text{interp}}(q)}{Q} \tag{7}$$

Table 4. Confuse matrix, which is often used to evaluate CNN's performance.

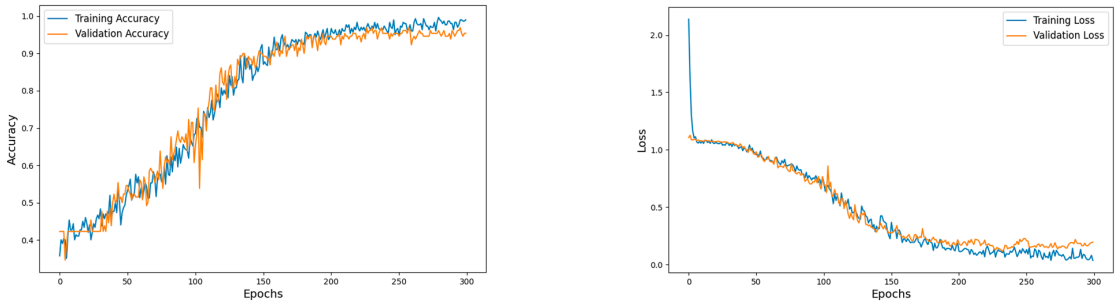
		Real	
		True	**False**
Predicted	True	Tp (True Positive)	Fp (False Positive)
	False	Fn (False Negative)	Tn (True Negative)

3. Results

This section analyzes the classification performance of the developed CNN model for social–emotional music. Figure 5a,b illustrate the model's performance during the training and validation processes, where the blue line represents training accuracy, and the orange line represents validation accuracy. According to the experimental results, the model achieved 90% training accuracy at epoch 180 and continued to converge between epochs 180 and 300, reaching the highest validation accuracy of 92.12% at epoch 280. Simultaneously, the validation loss remained stable during this period, indicating that the model did not exhibit significant overfitting. The final social–emotional music confusion matrix, presented in Figure 6, demonstrates that most social–emotional music samples were accurately classified into their respective categories.

This study employs repeated validation with ten iterations to ensure the stability and reliability of the model. After discarding the two worst-performing results, the average of the remaining results is taken as the final accuracy metric. This approach effectively reduces variations caused by weight initialization, stochastic batch training, and data partitioning, mitigating the uncertainty introduced by a single training instance in model evaluation. The experimental results are shown in Table 5 and indicate that the proposed CNN model achieves an average accuracy of 89.69%, with a maximum accuracy of 92.13%. Additionally, the model attains an average recall of 88.70% (maximum 90.06%), an average precision of 88.81% (maximum 90.74%), and an average mean average precision (mAP) of 93.95%

(maximum 94.72%). These results demonstrate the model's stability and effectiveness in social–emotional music classification tasks.



(**a**) (**b**)

Figure 5. Social–emotional CNN training process in 300 epochs. (**a**) SEM-Net accuracy training process. (**b**) SEM-Net loss function training process.

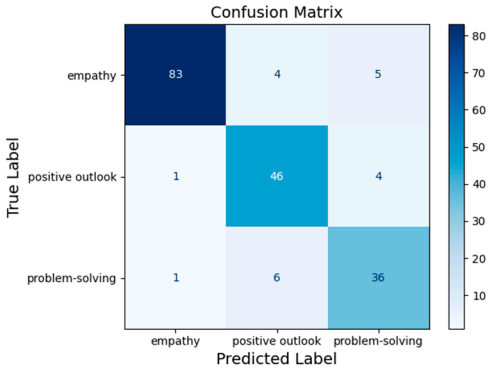


Figure 6. SEM-Net confuse matrix.

Table 5. Ten-fold cross-validation.

Round	Accuracy	Recall	Precision	mAP
Round 1	87.6344	87.4898	86.1046	89.4479
Round 2	89.3518	89.2071	87.3309	94.6216
Round 3	90.8602	89.2508	89.9715	94.5757
Round 4	90.3225	89.3749	88.31713	94.5399
Round 5	87.0967	85.9624	85.4076	95.2868
Round 6	89.7849	88.7694	89.4971	94.1361
Round 7	88.1720	86.8567	86.9572	92.9899
Round 8	87.6344	87.4898	87.4456	93.9839
Round 9	89.2473	88.6984	87.8575	94.7236
Round 10	92.1296	90.0116	90.7352	92.0842
Average	90.6878	88.7073	88.5140	93.9568

Furthermore, we analyze the performance of different proposed social–emotional CNN models before and after applying the social–emotional music preprocessing method. As shown in Table 6, the accuracy before preprocessing was only 42.31%, whereas the proposed method improved accuracy to 89.35%, representing an approximate 57% increase, demonstrating significant performance enhancement. Additionally, recall increased from 35.31% to 89.21%, precision increased from 35.28% to 87.31%, and mAP rose from 32.61% to 94.62%. These results indicate that the proposed method achieves substantial improvements

across all evaluation metrics, further validating the effectiveness of the social–emotional CNN model in social–emotional music classification tasks.

Table 6. SEM-Net classification comparison before and after pre-processing.

	Before	After
Accuracy	42.3107	89.3518
Recall	35.3129	89.2071
Precision	35.2817	87.3305
mAP	32.6175	94.6216

This study investigates how different feature extraction methods affect the performance of SEM-Net. The results of the ablation experiments are summarized in Table 7. Without feature extraction, the accuracy is only 52.78%. However, after applying various feature extraction techniques, classification accuracy improves significantly. Specifically, when using MFCC, Chroma, or Spectral Contrast individually, the accuracy reaches 80.60%, 81.94%, and 60.65%, respectively, demonstrating the importance of these features in music classification. Furthermore, combining multiple features leads to additional performance improvements. For instance, MFCC + Chroma achieves 76.85%, MFCC + Spectral Contrast reaches 77.31%, while Chroma + Spectral Contrast results in 64.52%. Notably, when integrating all three features (Chroma + Spectral Contrast + MFCC), the accuracy reaches 94.13%, representing a 41.35% improvement over the 54.78% accuracy without feature extraction.

Table 7. SEM-Net feature extraction comparison.

	Accuracy	Recall	Precision	mAP
Original	52.7817	33.3301	17.5909	33.3333
MFCC	80.5957	78.6009	78.3749	84.2512
Chroma	81.9440	82.4309	79.4573	88.4472
Spectral Contrast	60.6481	55.7368	60.5837	64.8835
MFCC + Chroma	76.8518	77.2238	74.9399	83.2052
MFCC + Spectral Contrast	77.3148	75.9861	75.1992	87.8583
Chroma + Spectral Contrast	64.5161	57.1965	64.5223	67.9755
Chroma + Spectral Contrast + MFCC	94.1296	93.0114	93.7352	92.0842

Table 8 presents the classification results for three types of social–emotional music. The model performed best on empathic perspective-taking, with an accuracy of 91.94% and mAP of 97.76%, followed by outlook and problem-solving, achieving accuracy above 90% and mAP over 95%. These results show that the model can effectively distinguish between different types of social–emotional music. While the performance in the others category was slightly lower, it still maintained reasonable accuracy, indicating the model's general classification ability.

Table 8. Three types of social–emotional music and other music classification result.

Type	Accuracy	Recall	Precision	mAP
Empathetic perspective-taking	91.9355	97.8261	93.7500	97.7639
Outlook	89.7849	90.1961	80.7018	96.1655
Problem-solving	90.8602	81.3953	87.5001	95.0258
Others	80.5555	83.9286	84.6847	84.2512

In the task of music classification, LSTM, Transformer, and BBNN represent different model strategies for feature processing. LSTM excels at handling sequential data and capturing long-term dependencies in music, making it suitable for modeling temporal features such as melody and rhythm [26]. On the other hand, Transformer utilizes a self-attention mechanism to simultaneously attend to different segments of an entire music piece, offering advantages in parallel computation and long-range dependency modeling [27]. It is particularly effective in capturing complex musical structures, though it generally demands higher computational resources and larger datasets. BBNN is a structurally flexible convolutional network that uses multi-branch and multi-scale feature extraction to fuse local and global audio features effectively [28], and CNN extracts local time–frequency features of music through convolutional layers [29]. BBNN and CNN often achieve efficient performance in classification tasks, mainly when applied to time–frequency representations like Mel-spectrograms. According to the data analysis in Table 9, the proposed model outperforms other models across all evaluation metrics, demonstrating superior classification performance. In the model comparison, the accuracy of LSTM and Transformer models was 42.86% and 38.33%, respectively, indicating their suboptimal performance in this study's application scenario. This may be attributed to LSTM's reliance on long-sequence temporal information and the Transformer's insufficient training data, which may have hindered its ability to learn musical features thoroughly. Additionally, the BBNN model achieved a relatively higher accuracy of 67.38%, yet it still fell short of the proposed method. The CNN + LSTM model achieved an accuracy of 53.59%, slightly outperforming LSTM alone but remaining below an ideal level. In contrast, the proposed method significantly outperformed all other models, achieving an accuracy of 94.13%, recall of 93.01%, precision of 93.74%, and mAP of 92.08%. To validate the superiority of the proposed approach, k-fold cross-validation (k = 10) was conducted, ensuring fairness in comparison by applying the same hyperparameter settings across different models. Furthermore, a statistical significance test was performed on the results, revealing that the proposed method achieved a p-value < 0.05 compared to other models, indicating that its advantage is statistically significant. These findings confirm the effectiveness of the proposed SEM-Net model in this application domain, demonstrating its suitability for social–emotional music classification compared to traditional models. Among all the models, the Transformer model has the highest number of parameters, reaching 350,274, which reflects the high computational resource demand caused by its multi-head attention mechanism and deep stacking architecture. In contrast, the LSTM model has the fewest parameters, only 145,411, indicating a more straightforward structure suitable for deployment in resource-constrained embedded systems. The SEM-Net model has 251,091 parameters, which is higher than that of LSTM (145,411) and BBNN (206,714) but lower than CNN + LSTM (267,500) and Transformer (350,274). SEM-Net maintains a reasonable level of parameter efficiency while still delivering strong performance without introducing excessive computational overhead.

Table 9. Comparison with different classification models.

Model	Accuracy	Recall	Precision	mAP	Parameter Quantity
LSTM [26]	42.8601	35.5647	40.0017	42.4727	145,411
Transformer [27]	38.3347	27.7685	51.4485	32.3627	350,274
BBNN [28]	67.7438	67.3816	68.7674	73.5882	206,714
CNN + LSTM [29]	53.8759	53.5971	61.9795	68.0274	267,500
SEM-Net (this study)	94.1296	93.0114	93.7352	92.0842	521,091

4. Discussion

This preliminary investigation aims to contribute to the growing literature on enhancing social–emotional learning for individuals with special needs through an innovative AI-based music classification model. The strong performance of SEM-Net is closely linked to the integration of appropriate data preprocessing techniques and complementary audio feature extraction methods. Specifically, time stretching, pitch shifting, and noise augmentation introduced controlled variability into the input data, which enhanced the model's ability to generalize across diverse musical expressions. Three feature extraction methods—MFCC, Chroma, and Spectral Contrast—were selected for their ability to capture different yet complementary aspects of musical content. MFCC captures timbral characteristics, Chroma reflects pitch and harmonic relationships, and Spectral Contrast emphasizes variations in energy across frequency bands. As shown in Table 7, although each feature individually achieved reasonable accuracy, their combination significantly improved overall performance, reaching 94.13%. This demonstrates that these features complement each other effectively and enable SEM-Net to capture the nuanced patterns in social–emotional music better. Thus, this study has three contributions: (1) Proposed an innovative social–emotional music classification model," SEM-Net'', to help those people who need special assistance. (2) Through the social–emotional pre-processing and feature extraction method, SEM-Net can achieve the highest accuracy of 94.13%. (3) Compared with traditional models such as LSTM or Transformer, SEM-Net achieves at least 20% higher classification accuracy. Although its parameter count is slightly higher than parts of traditional models, SEM-Net demonstrates a more favorable balance between model complexity and performance. This demonstrates the effectiveness of the proposed method in accurately categorizing social–emotional music.

However, this study has certain limitations. The first limitation lies in the dataset used for training and evaluation. While the model performs well within the scope of the selected dataset, it does not comprehensively represent all relevant music genres or styles that may influence social–emotional learning. As a result, its applicability may be restricted to specific scenarios, and further work is needed to expand and diversify the dataset to improve generalizability. The second limitation concerns the scalability of the proposed model. Although the model demonstrates strong classification capabilities, its ability to process large-scale music databases in real-world applications remains uncertain. Future research should enhance computational efficiency and optimize the model for large-scale, real-time music classification tasks. This could involve exploring self-supervised learning techniques, data augmentation strategies, and more efficient deep learning architectures to improve performance and efficiency.

Future studies should investigate how social–emotional music classification can be integrated into personalized learning environments or therapeutic applications. One possible direction is to develop an adaptive AI system that can dynamically adjust to the emotional needs of individuals by recommending suitable social–emotional music. Additionally, incorporating multimodal data (e.g., physiological signals, facial expressions, or behavioral responses) could further refine the model's ability to interpret emotional states and provide more personalized feedback. In summary, while this study presents a promising approach to social–emotional music classification, further research is necessary to address dataset limitations, enhance model scalability, and explore its integration into real-world applications. Future advancements in deep learning, affective computing, and personalized AI systems will play a crucial role in enhancing the practical impact of AI-driven social–emotional learning tools.

5. Conclusions

This study successfully developed an AI-based social–emotional music classification model by integrating SEL concepts and musical literacy theory. SEM-Net features a multi-feature fusion architecture that combines MFCC, Chroma, and Spectral Contrast to capture complementary aspects of musical emotion. In addition, data augmentation techniques were applied to enhance model generalization. Experimental results demonstrated that SEM-Net achieved an accuracy of 94.13%, significantly outperforming existing models by at least 20%. Despite having a moderately higher parameter count than simpler models, SEM-Net maintains a good balance between efficiency and performance. The findings suggest that classical music—characterized by expressive melodies, diverse harmonies, and structured rhythms—can effectively foster individuals' positive outlook, empathy, and problem-solving skills, enhancing emotional resilience and social competence.

Author Contributions: Conceptualization, Y.-C.C. and P.-C.C.; Data curation, S.-K.C.; Formal analysis, S.-K.C.; Funding acquisition, S.-L.C. and Y.-C.C.; Methodology, C.-Y.C., K.-K.Y., Y.-C.P. and C.-H.T.; Resources, Y.-C.C.; Software, Y.-J.L., C.-Y.C., K.-K.Y., Y.-C.P. and C.-H.T.; Validation, Y.-J.L.; Visualization, Yuan-Jin Lin, C.-Y.C., K.-K.Y., Y.-C.P. and C.-H.T.; Writing—original draft, Y.-J.L.; Writing—review and editing, K.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science and Technology Council (NSTC), Taiwan, under grant numbers NSTC-112-2410-H-033-014, NSTC-112-2221-E-033-049-MY3, NSTC-113-2622-E-033-001 and NSTC-113-2410-H-033-021-MY2, and the National Chip Implementation Center, Taiwan.

Institutional Review Board Statement: The classical music databases used by this institute are all open resources and do not contain any institutional review board statement.

Informed Consent Statement: All authors agree to publish this study.

Data Availability Statement: The relevant research data for this study cannot be provided due to other project needs.

Acknowledgments: The authors thank the Applied Electrodynamics Laboratory (Department of Physics, National Taiwan University) for supporting the microwave calibration kit and components.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baurain, C.; Nader-Grosbois, N. Theory of Mind, Socio-Emotional Problem-Solving, Socio-Emotional Regulation in Children with Intellectual Disability and in Typically Developing Children. *J. Autism Dev. Disord.* **2013**, *43*, 1080–1097. [CrossRef] [PubMed]
2. Davidson, R.J.; Begley, S. *The Emotional Life of Your Brain: How Its Unique Patterns Affect the Way You Think, Feel, and Live—And How You Can Change Them*; Avery Publishing Group: New York, NY, USA, 2012.
3. Nguyen, W.; Ownsworth, T.; Nicol, C.; Zimmerman, D. How I See and Feel About Myself: Domain-Specific Self-Concept and Self-Esteem in Autistic Adults. *Front. Psychol.* **2020**, *11*, 913. [CrossRef] [PubMed]
4. Piers, L.; Duquette, C.A. Facilitating Academic and Mental Health Resilience in Students with a Learning Disability. *Except. Educ. Int.* **2016**, *26*, 21–41. [CrossRef]
5. Bouton, B. Empathy Research and Teacher Preparation: Benefits and Obstacles. *SRATE J.* **2016**, *25*, 16–25.
6. Shamay-Tsoory, S.G. The Neural Bases for Empathy. *Neuroscientist* **2011**, *17*, 18–24. [CrossRef]
7. Levinson, J. Musical Literacy. *J. Aesthetic Educ.* **1990**, *24*, 17–30. [CrossRef]
8. Nwokenna, E.N.; Sewagegn, A.A.; Falade, T.A. Effect of educational music intervention on emotion regulation skills of first-year university music education students. *Medicine* **2022**, *101*, e32041. [CrossRef]
9. Peters, V.; Bissonnette, J.; Nadeau, D.; Gauthier-Légaré, A.; Noël, M.-A. The impact of musicking on emotion regulation: A systematic review and meta-analysis. *Psychol. Music* **2024**, *52*, 548–568. [CrossRef]
10. Watts, S.H.; Piña, K. Bibliotherapy and Social and Emotional Learning in the Elementary Music Setting. *J. Gen. Music Educ.* **2023**, *36*, 5–12. [CrossRef]

11. Eren, B. Use of Music in Special Education and Application Examples from Turkey. *Procedia—Soc. Behav. Sci.* **2014**, *116*, 2593–2597. [CrossRef]
12. Jorgensen, E.R. Western Classical Music and General Education. *Philos. Music Educ. Rev.* **2003**, *11*, 130–140. [CrossRef]
13. Atallah, H.E.; Lopez-Paniagua, D.; Rudy, J.W.; O'Reilly, R.C. Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat. Neurosci.* **2007**, *10*, 126–131. [CrossRef] [PubMed]
14. Jobst, B.C.; Gonzalez-Martinez, J.; Isnard, J.; Kahane, P.; Lacuey, N.; Lahtoo, S.D.; Nguyen, D.K.; Wu, C.; Lado, F. The Insula and Its Epilepsies. *Epilepsy Curr.* **2019**, *19*, 11–21. [CrossRef] [PubMed]
15. Mitterschiffthaler, M.T.; Fu, C.H.Y.; Dalton, J.A.; Andrew, C.M.; Williams, S.C.R. A functional MRI study of happy and sad affective states induced by classical music. *Hum. Brain Mapp.* **2007**, *28*, 1150–1162. [CrossRef]
16. Baumgartner, T.; Esslen, M.; Jäncke, L. From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *Int. J. Psychophysiol.* **2006**, *60*, 34–43. [CrossRef]
17. Rosner, A.; Kostek, B. Automatic music genre classification based on musical instrument track separation. *J. Intell. Inf. Syst.* **2017**, *50*, 363–384. [CrossRef]
18. Anglade, A.; Benetos, E.; Mauch, M.; Dixon, S. Improving Music Genre Classification Using Automatically Induced Harmony Rules. *J. New Music Res.* **2010**, *39*, 349–361. [CrossRef]
19. Pérez-Sancho, C.; Rizo, D.; Iñesta, J.M.; de León, P.J.P.; Kersten, S.; Ramirez, R. Genre classification of music by tonal harmony; IOS Press. *Intell. Data Anal.* **2010**, *14*, 533–545. [CrossRef]
20. Fonseca, E.; Favory, X.; Pons, J.; Font, F.; Serra, X. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 829–852. [CrossRef]
21. Joshy, A.A.; Rajan, R. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 1147–1157. [CrossRef]
22. Zalkow, F.; Müller, M. CTC-Based Learning of Chroma Features for Score–Audio Music Retrieval. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2957–2971. [CrossRef]
23. Rí, F.A.D.; Ciardi, F.C.; Conci, N. Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks. *IEEE Access* **2023**, *11*, 116638–116649. [CrossRef]
24. Panda, R.; Malheiro, R.; Paiva, R.P. Audio Features for Music Emotion Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2023**, *14*, 68–88. [CrossRef]
25. Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review. *Electronics* **2022**, *11*, 3795. [CrossRef]
26. Zhang, S.; Liu, Y.; Zhou, M. Graph Neural Network and LSTM Integration for Enhanced Multi-Label Style Classification of Piano Sonatas. *Sensors* **2025**, *25*, 666. [CrossRef]
27. Wang, J.; Li, S.; Sung, Y. Deformer: Denoising Transformer for Improved Audio Music Genre Classification. *Appl. Sci.* **2023**, *13*, 12673. [CrossRef]
28. Liu, C.; Feng, L.; Liu, G.; Wang, H.; Liu, S. Bottom-up broadcast neural network for music genre classification. *Multimed. Tools Appl.* **2021**, *80*, 7313–7331. [CrossRef]
29. Gupta, R.; Ashish, S.; Shekhar, H.; Dominic, D.S. Music Genre Classification Using CNN and RNN-LSTM. In *Micro-Electronics and Telecommunication Engineering*; Sharma, D.K., Peng, S.-L., Sharma, R., Zaitsev, D.A., Eds.; Springer Nature: Singapore, 2022; pp. 729–745. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.